Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240536

From EHR to Machine Learning: A Preliminary Report on an Ingestion Pipeline Based on JSON-LD

Giulia Lucrezia BARONI[a] , Vincenzo DELLA MEA [a,1] and Gian Luca FORESTI[a]
[a] *Dept. of Mathematics, Computer Science and Physics, University of Udine, Italy*
ORCiD ID: Vincenzo Della Mea https://orcid.org/0000-0002-0144-3802

Abstract. In this paper, we present the preliminary experiments for the development of an ingestion mechanism to move data from Electronic Health Records to machine learning processes, based on the concept of Linked Data and the JSON-LD format.

Keywords. EHR, Machine Learning, Linked Data

1. Introduction

In the last few years, the availability of novel machine learning methods fostered advances in various areas of clinical practice, starting from those involving biomages and biosignals. However, of particular interest is also machine learning applied to health data extracted from Electronic Health Records (EHR), at least is sectors where data collection happens mostly in electronic form. EHRs are commonly supported by relational databases. Furthermore, part of the data can be coded using terminologies and classifications like ICD, SNOMED-CT, LOINC, and other. Finally, sometimes EHRs can communicate with other health information systems by means of the available health informatics standards, like HL7, CDA and FHIR. When experimenting and implementing machine learning (ML) within health information systems, the first step is to build a preprocessing pipeline able to extract the needed data from the EHR in a suitable form for model ingestion. We report on preliminary experiments carried out in the framework of a Rare Diseases project.

2. Methods

We based our analysis on the Epidemiological Information System of the Region Friuli Venezia Giulia, Italy. By examining the underlying database, consisting of 23 modules for a total of 179 tables, we found that there were at least 9 dictionaries potentially related to internationally defined artifacts (diseases, procedures, drugs, laboratory exams, etc) and other 11 locally defined (geographical locations, health structures, etc).

Notwithstanding the abundant data currently digitized in EHRs, recent evidence demonstrates that only a fraction of such data – on average 27 variables, and mostly from

[1] Corresponding Author: Vincenzo Della Mea, E-mail: vincenzo.dellamea@uniud.it.

a single system- is commonly exploited in machine learning [2]. One reason is the lack of technical or semantic interoperability. One common format is CSV, being it easily manageable with all machine learning frameworks and libraries, but it lacks the capability of fully representing the richness of coded data. Having this in mind, we wanted to identify a possible path to simplify the ingestion of data, taking also in account the richness given by dictionaries, terminologies and classifications, which is often lost when data is converted to CSV or equivalent formats.

3. Results

Our proposal relies on the Linked Data concept, and in particular on the JSON-LD format. JSON-LD provides a lightweight framework for linked data, that is, data interconnected with other data. In our case, EHR coded data should be connected with biomedical dictionaries in an explicit way, otherwise lost with simpler formats like CSV. JSON-LD is centered around the 'context', which allows to relate one or more variables to concepts specified in ontologies. In our case, it is straightforward to use this approach to link coded data with the source of codes, i.e., the ontology or classification. At the time of training or inference, the ML system has the possibility to trace back the origin of coded data through the JSON-LD context without the need for hardcoding it.

4. Discussion

An approach that preserves the linkage between coded data and their coding infrastructure might help to respect the richness of health data and better implement machine learning, in particular specific loss functions exploiting the hierarchical structure of ontologies [4]. JSON-LD might represent a way to reach this aim, yet being a lightweight format with less syntactic and semantic complexities as FHIR or CDA.

Acknowledgments

This work is supported by the Friuli Venezia Giulia Region project F53C22001780002 *"Automatic evaluation of diagnostic images using AI by harnessing the ability of algorithms to process, identify and classify any anomalies or lesions to support operators in the diagnostic field"*.

References

[1] Casey A. Health Informatics Standards. In: Hannah, K., Hussey, P., Kennedy, M., Ball, M. (eds) Introduction to Nursing Informatics. Health Informatics. Springer, London, 2015.

[2] Goldstein, BA, Navar AM, Pencina MJ, Ioannidis, JPA Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J. Am. Med. Inform. Assoc. 2017;24:198–208.

[3] Kellogg G, Champin PA, Longley D. JSON-LD 1.1 – A JSON-based Serialization for Linked Data (W3C Working Draft). [Technical Report] W3C. 2019. ⟨hal-02141614v1⟩

[4] Amigo E, Delgado A. Evaluating Extreme Hierarchical Multi-label Classification. In Proc. of the 60th Annual Meeting of the ACL, pages 5809–5819, Dublin, Ireland.